Cell Reports, Volume *32*

Supplemental Information

Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts

Jeffrey S. McLean, Batbileg Bor, Kristopher A. Kerns, Quanhui Liu, Thao T. To, Lindsey Solden, Erik L. Hendrickson, Kelly Wrighton, Wenyuan Shi, and Xuesong He

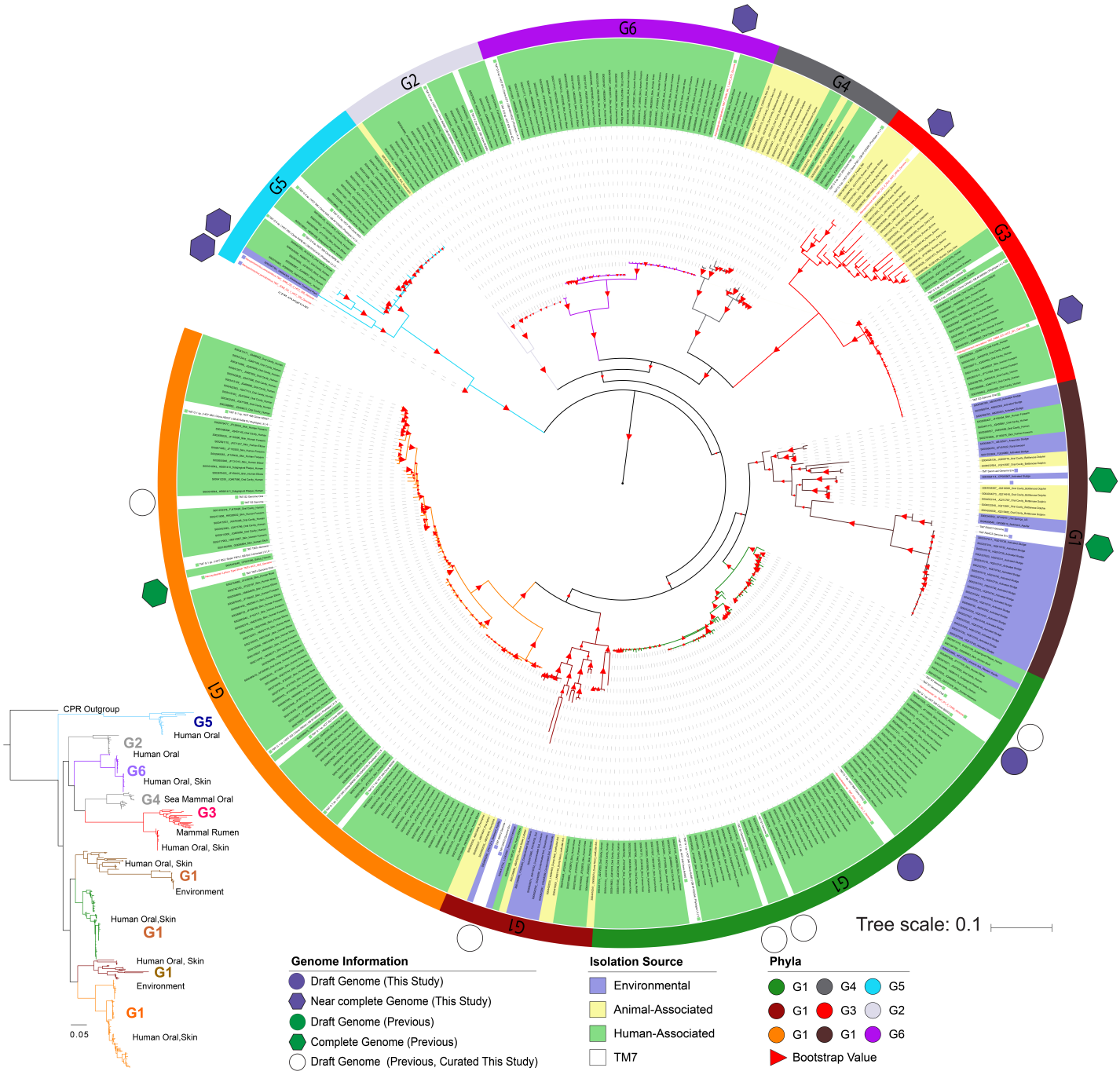


Figure S1. Phylogeny of Phylum Saccharibacteria based on SSU rRNA. Related to Figure1. (see also Supplemental Data File 1. Newick Tree). Full length sequences of publically available genomes and those from this study along with the HOMD representative sequences were searched against the Silva Database (nearest neighbors references (n-20; 89% cutoff) and extracted for each sequence following alignment and trimming. The resulting 400 sequences in the SILVA alignment were masked with >10% gaps removed. Trees were inferred by Maximum likelihood (RAXML) with 100 bootstrap resamplings and metadata included to highlight sources. CPR genomes from SR-1, Kazan and Berkelbacteria were used as outgroups.

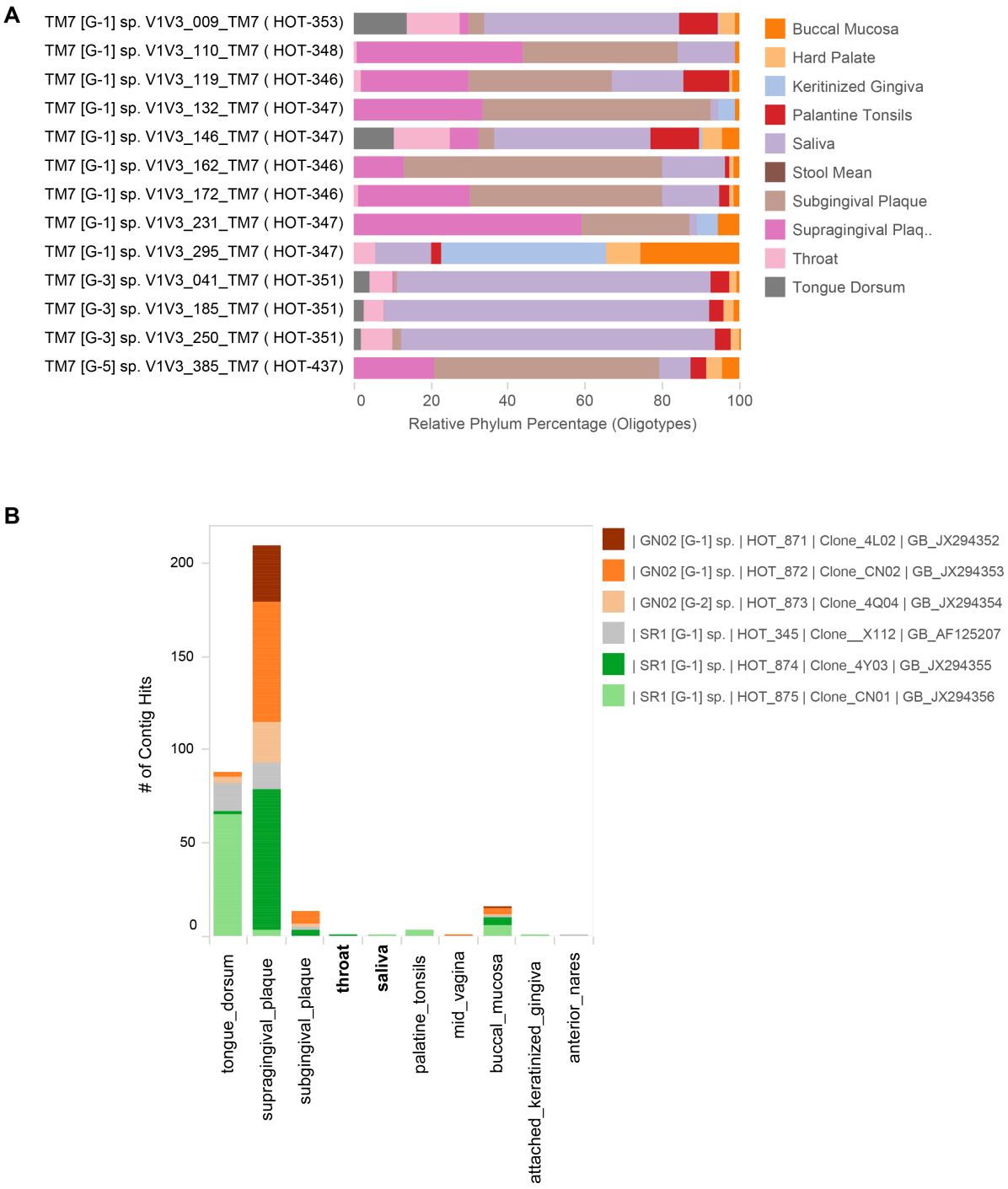


Figure S2. Distribution of CPR Phylum across body sites from the HMP datasets. Related to Figure 4. (**A**) curation of available oligotyping data of HMP data to show the relative percentages of the phylum in each body site derived using 16S rRNA V1-V3 datasets (~250 nucleotides) from Eren et al. 2014) . (**B**) number of hits to 16S rRNA genes from each GN02 and SR-1 groups in 1129 HMP body site assemblies (cutoff ; > 299 bp and > 79% pairwise identity).

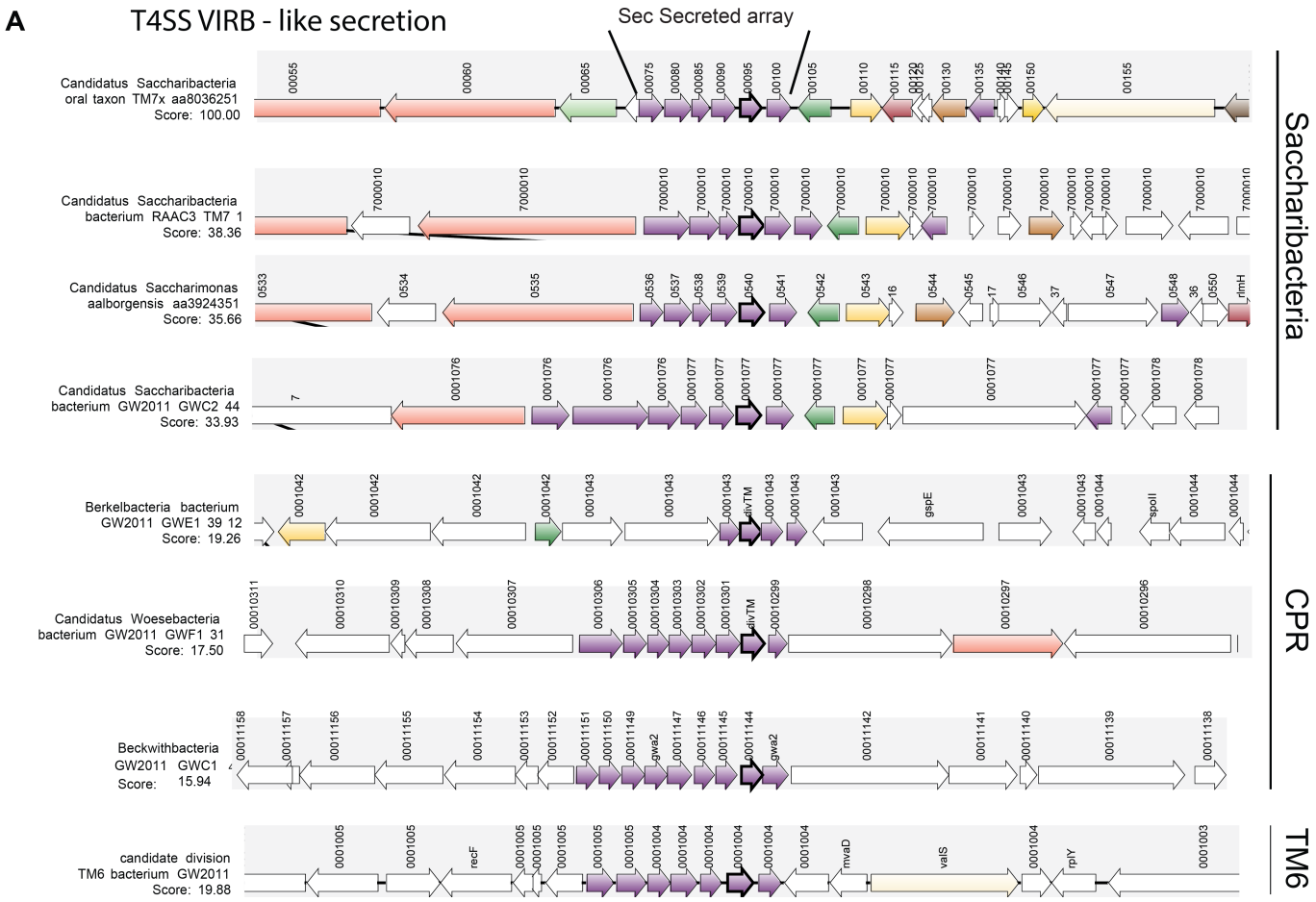


Figure S3. T4SS secretion system shared among genomes. Related to Figure 5. A novel region that has an array of small hypothetical proteins (4-6) all containing N-terminal signal peptides were found within the syntenic region of the G1 group. This unique array of proteins, also present in G3, G5 and G5 genomes (**Figure 5)** was maintained in other Saccharibacteria, CPR and the reduced genome amoebae symbiont belonging to the lineage TM6 ((29) current proposed Phylum name for TM6 is Dependentiae) shown here. Variations are present in the number of proteins in this array. The closest homology of the proteins in the neighborhood of in this array to any known sequence indicates it is related to VIRB2 type IV secretion system and may likely be required for parasitism of their bacterial hosts or amoeba in the case of TM6.

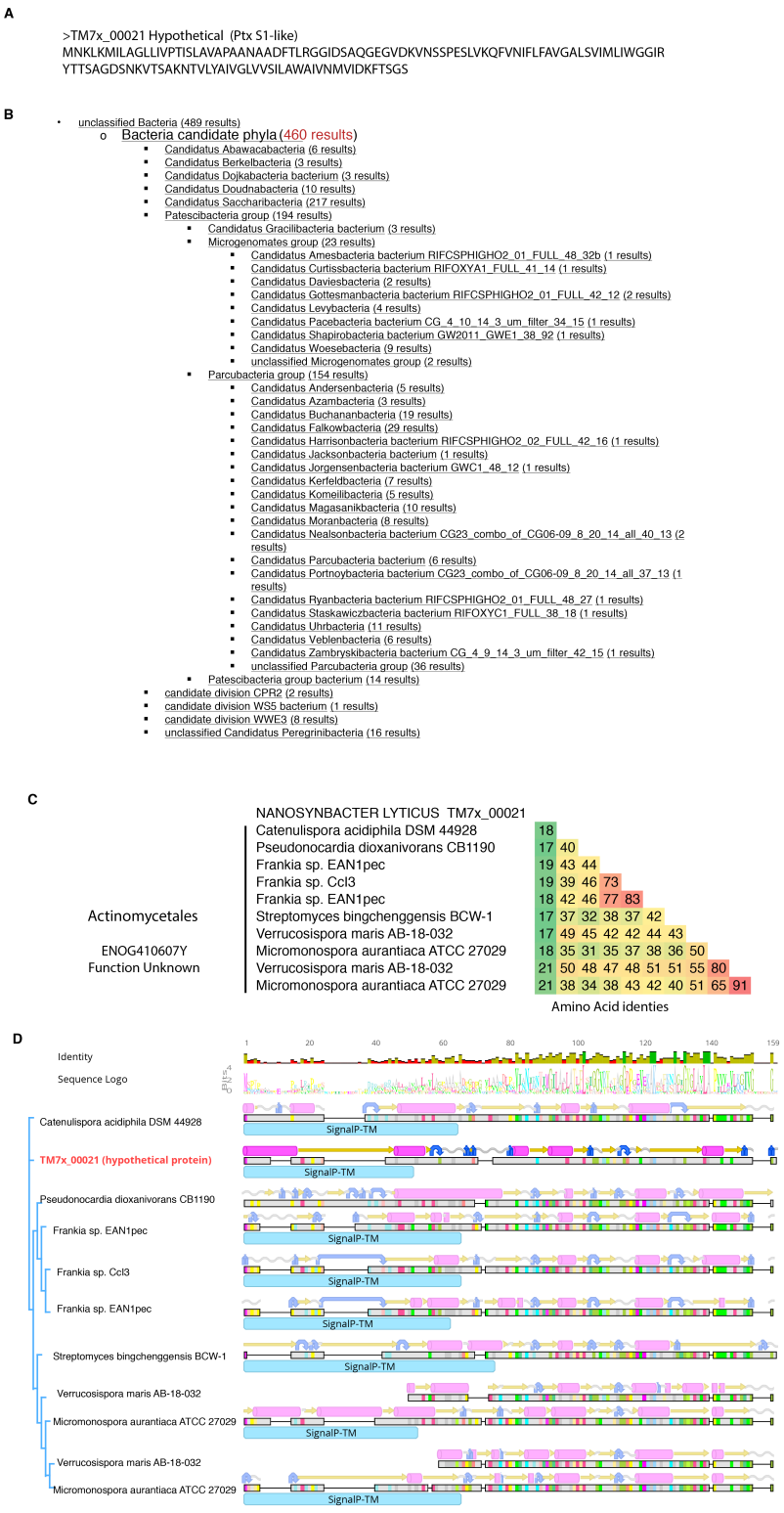


Figure S4. **Distribution and orthologs of the T4SS related small secreted hypothetical protein from *N. lyticus* strain Tm7x (locus tag TM7x_00021)**. **Related to Figure 4**. (**A**) amino acid sequence of the small hypothetical protein in TM7x, TM7x_0002, that was most similar to the Ptx S1 toxin protein in *B. pertussis*. (**B**) Blast searches in Uniprot database identified this protein in nearly all other Saccharibacteria and across the bacterial candidate phyla, mainly CPR. Number in parentheses indicates the number detected for each genome/assembly. (**C**) Amino acid identities in orthologous gene hits outside of the candidate phyla, grouped in the Eggnog database under entry ENOG410607Y, are shown as a matrix. Amino acid identities increase from green to red with red as the highest identities. All of these top hits were found in the Actinomycetales order to which TM7x basibiont belongs. (**D**) Gapped alignment of TM7x_00021 with the proteins in C.

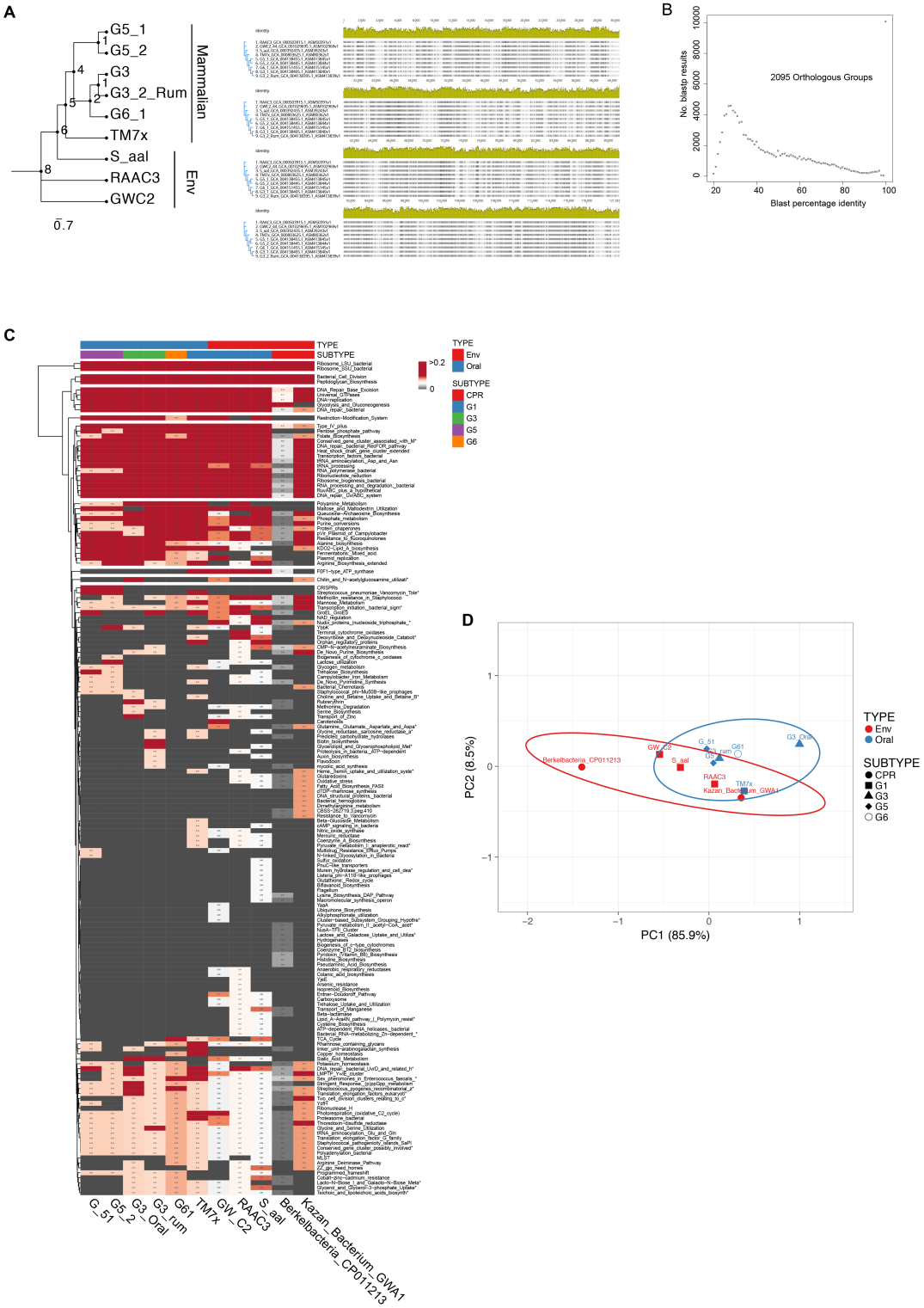
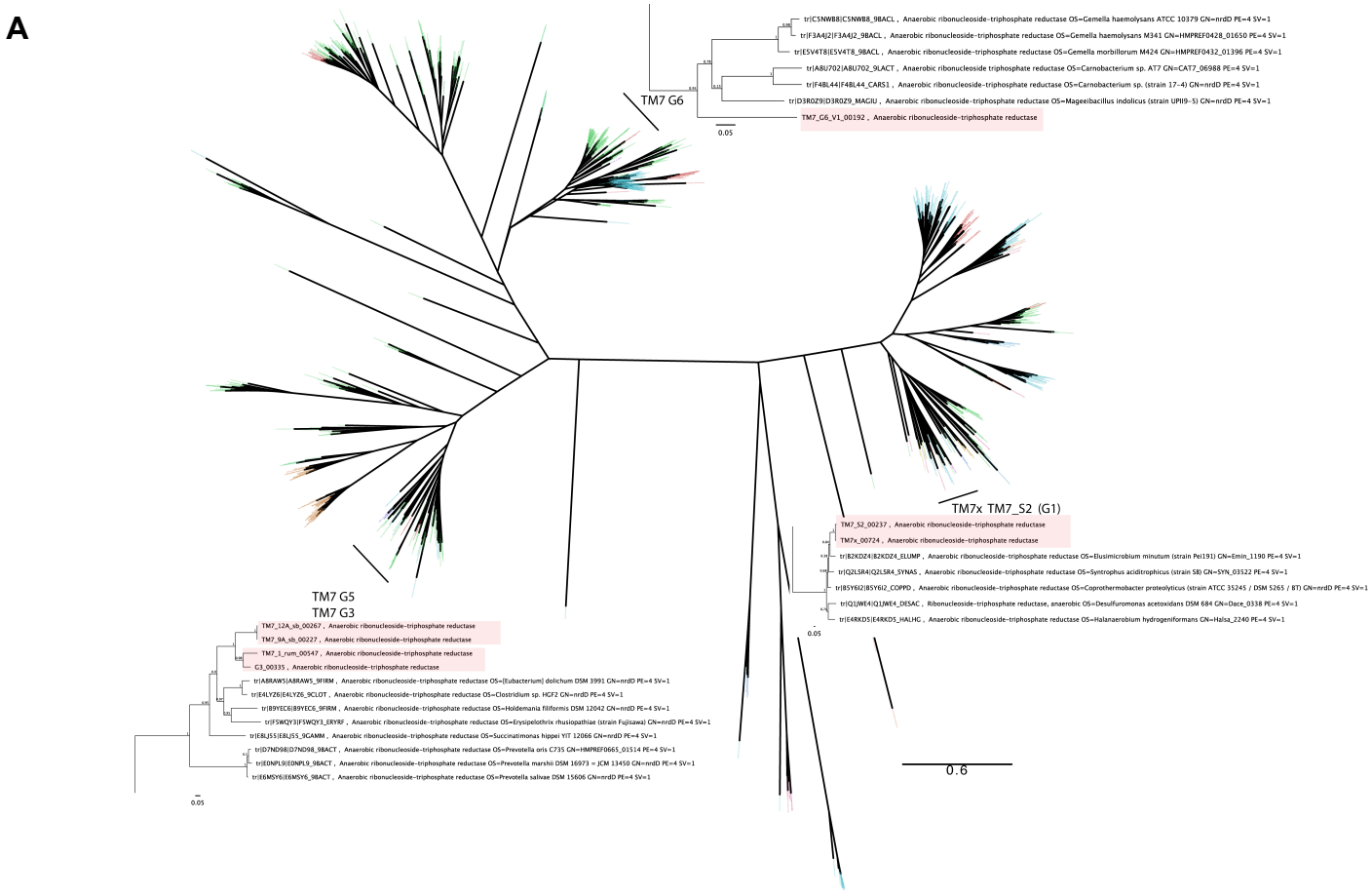


Figure S5. Core gene alignment tree and comparative analyses of predicted gene functions within the Saccharibacteria phylum. Related to Figure 5. (**A**) Core gene alignment and inferred tree of the 211 concatenated shared genes within four complete and five near complete Saccharibacteria genomes (RAxML GTR-CAT, 1000 bootstrap) agrees with the G3, G5, G6 groups being the most distant and earlier branching of the G1 clades showing the environmental derived metagenome assemblies are closer to the root of the tree ,in agreement with the 16S rRNA gene in **Figure 1C and 1D**. (B) BLAST percentage identity and number of pairwise BLAST results highlights the overall low percentage identities across the compared genomes. (**C**) Row clustered heatmap of percentage of genes in the genome for predicted functions within each assembled genome. Genome type (column) and functions (rows). Rows are centered and normalized. Select CPR that have been used as outgroups previously are also included. Shared as well as unique functions defining the groups are visible. (**D**) Principal component analysis of data in (c) for selected genomes with high completeness showing functional clustering by groups (i.e. oral G1 and environmental G1).


A
TM7 G6
TM7 G5
TM7 G3
TM7x TM7_S2 (G1)
0.6
0.05

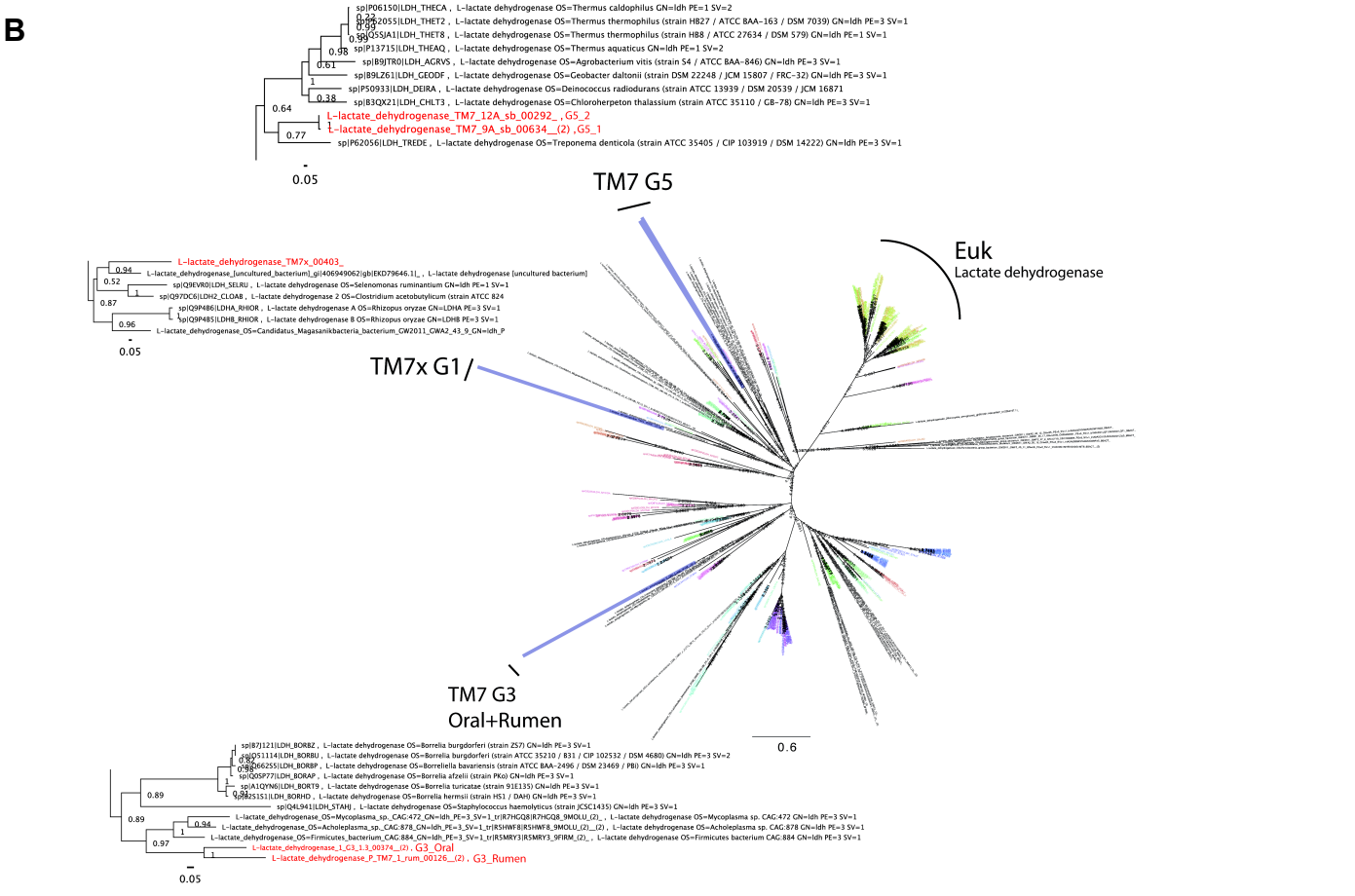

B
TM7 G5
Euk
Lactate dehydrogenase
TM7x G1
TM7 G3
Oral+Rumen
0.6
0.05

Figure S6. Maximum likelihood phylogenetic tree of proteins acquired in human associated groups. Related to Figure 6. (**A**) The gene encoding the protein NrdD was identified as being unique among the oral genomes that was not present in the environmental genomes. The taxonomic best hits (**Figure 6**) and the protein trees shown here indicated that divergent phylogenetic relatedness between the acquired proteins supporting independent acquisition from different bacteria. This anaerobic version of the ribonucleoside-triphosphate reductase would enable function under host conditions in the oral cavity. In addition, the G3 (oral and rumen) and G5 groups have *nrdG* gene -which enables activation of anaerobic ribonucleoside-triphosphate reductase under anaerobic conditions by generation of an organic free radical, using S-adenosylmethionine and reduced flavodoxin as co-substrates to produce 5'-deoxy-adenosine. Red highlights indicate the Saccharibacteria proteins from this study. **(B)** The Ldh protein was identified as being a unique among the oral genomes that was not present in the environmental genomes. The taxonomic best hits (**Figure 6**) and the gene trees for this gene indicate the divergent phylogenetic relatedness between the acquired genes amongst the different groups, supporting independent acquisition from different bacteria. Red text indicates the Saccharibacteria proteins from this study.

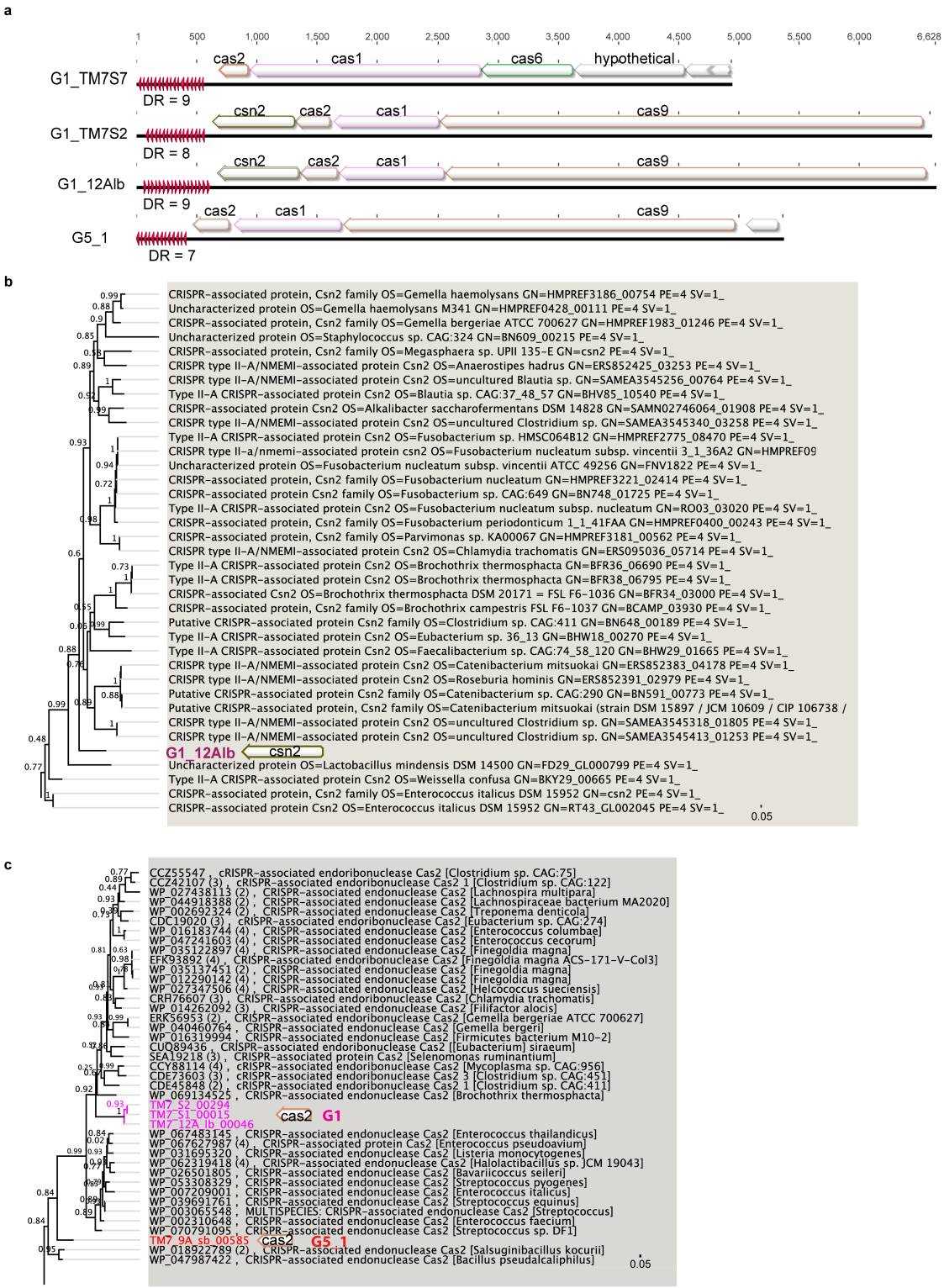


Figure S7. Diversity of genes, repeats, spacers and operon structure of CRISPR regions in Saccharibacteria. Related to Figure 6. (A) CRISPR regions for Saccharibacteria were not reported in previous published genomes and they were thought to lack this phage immunity mechanism. This study found CRISPR systems in several oral genomes, G1 (S7, S2, and 12Alb) and the G5. So far, no CRISPR have been detected in environmental representatives. No overlaps in the direct repeat sequences or the spacer sequences were found between genomes. Uniquely, a *cas6* gene was found within the G1_TM7S7 which was also missing the cas9 gene. The variability between G1 and G5 as well as some variability within the G1 also support independent acquisition. **(B)** tree for the *csn2* proteins unique to the G1 12Alb genome that clusters with host associated bacteria. **C)** tree for the G1 and G5 *C*as2 proteins, supporting independent acquisition from unrelated bacterial groups.